Rank parameters for Bland–Altman plots

Roger B. Newson

May 21, 2018

1 Introduction

Bland–Altman plots were introduced by Altman and Bland (1983)[1] and popularized by Bland and Altman (1986)[2]. Given N bivariate data points (A_i, B_i), expressed in the same units and assumed to be two alternative quantities used to measure the same thing (such as exam marks of the same students allocated by Lecturers A and B), a Bland–Altman plot is derived by rotating the scatter plot of the A_i (on the vertical axis) and the B_i (on the horizontal axis) through 45 degrees by transforming the bivariate column vectors $(A_i, B_i)^T$ with the rotation matrix

$$R = \begin{bmatrix} 1 & -1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{1}$$

Alternatively, in simpler language, it is defined by plotting the differences $A_i - B_i$ on the vertical axis against the means $(A_i + B_i)/2$ on the horizontal axis.

The Bland–Altman plot is a good graphical summary of the paired data points (A_i, B_i). However, we might prefer also to have confidence intervals for parameters of the Bland–Altman plot, in order to be able to make quantitative statements on how the A_i and the corresponding B_i agree and/or disagree with each other. In Section 4.7 of van Belle (2008)[10], it is argued (following Lin (1989)[5]) that the three principal components of disagreement are discordance (also known as "imprecision"), bias, and scale differential. Suppose that the (A_i, B_i) are sampled from a common bivariate distribution, with means μ_a and μ_B, standard deviations σ_A and σ_B, and correlation coefficient ρ_{AB}. Disagreement can then be decomposed into these three components by the formula

$$\frac{E[(A-B)^2]}{2\sigma_A\sigma_B} = (1-\rho_{AB}) + \frac{(\mu_A-\mu_B)^2}{2\sigma_A\sigma_B} + \frac{(\sigma_A-\sigma_B)^2}{2\sigma_A\sigma_B}, \tag{2}$$

where the left–hand side measures general disagreement, the first term in the right–hand side measures discordance, the second term in the right–hand side measures squared bias, and the third term in the right-hand side measures squared scale differential.

It will be argued here that all three principal components of disagreement are better measured using rank parameters than using regression parameters, although the regression parameters may be good proxies for the rank parameters if the (A_i, B_i) are sampled from a bivariate Normal distribution. This is because the rank parameters are less influenced by outlying data points, and also easier to interpret in words as measuring the appropriate type of disagreement. The rank parameters used will be Kendall's τ_a, estimated using the methods of Newson (2006a)[6], and median differences, estimated using the methods of Newson (2006b)[7].

1.1 Example dataset: Double marking of exam candidates

We will demonstrate the methods using an example dataset with one observation for each of 179 candidates in a medical school statistics examination, and data on marks awarded by 2 academics, who we will call "Lecturer A" and "Lecturer B". Lecturer A was the more experienced of the two, and Lecturer B was marking exams for the first time (in the course of an all-night session on heavy doses of coffee). There were 5 questions posed in the exam, of which students were asked to attempt 4. Each academic awarded each student a total mark, equal to the sum of the student's marks on the best 4 questions attempted. (It was not unknown for a student to attempt all 5 questions.) 3 of the candidates did not sit the exam, leaving 176 students who could be marked. Students were finally awarded the mean of the marks awarded by the two academics. The results are analysed using the Stata statistical software[9], particularly the add–on package `somersd`[6][7], and two other add–on packages `scsomersd` and `rcentile` (Newson, 2014)[8], which depend on `somersd`.

Figure 1 gives a scatter plot of the mark awarded by Lecturer A against the mark awarded by Lecturer B. Figure 2 gives a Bland–Altman plot of the difference between the mark awarded by Lecturer A and the mark

Figure 1: Scatter plot of marks awarded by Lecturers A and B.

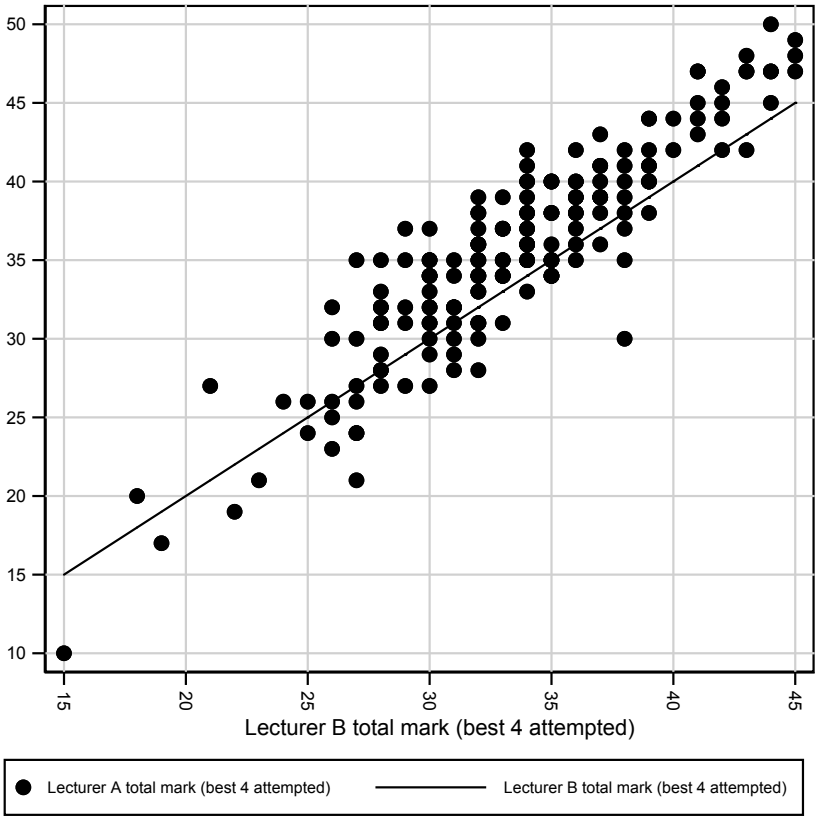
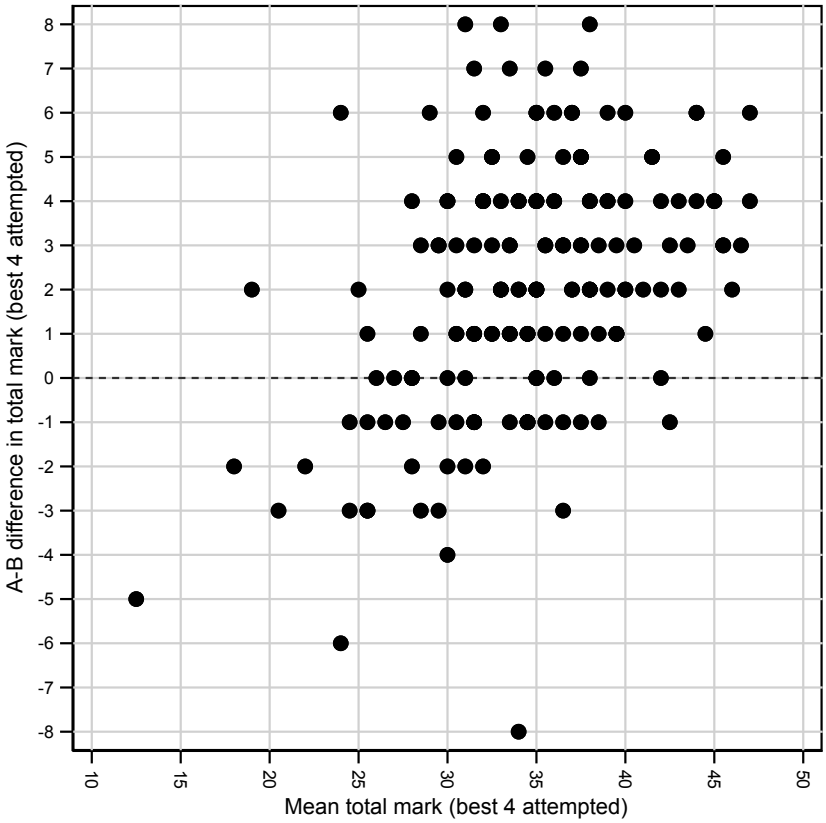


Figure 2: Bland–Altman plot of marks awarded by Lecturers A and B.



awarded by Lecturer B against the mean of the two marks (eventually awarded to the student). The vertical–axis reference line at zero in Figure 2 corresponds to the diagonal line of equality in Figure 1. Note that the Bland–Altman plot is more efficient in using space, as there is no wasted space in the top left and bottom right parts of the plot region. This allows us to view the vertical axis of the Bland–Altman plot in greater detail. In particular, we can see that the differences are integer, instead of being continuous.

2 Discordance parameters

We assume a population of indexed individuals, with 2 scalar variables X and Y defined for each individual, and a sampling scheme for sampling pairs of individuals (indexed as i and j) at random from that population. Kendall's τ_a of X and Y is then defined as

$$\tau(X,Y) = E\left[\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)\right], \tag{3}$$

where $E(\cdot)$ denotes expectation, and sign(x) is 1 if $x > 0$, -1 if $x < 0$, and 0 if $x = 0$. Alternatively, if we define the "cordance sign"

$$\text{csign}(x_i, y_i, x_j, y_j) = \text{sign}(x_i - x_j)\text{sign}(y_i - y_j), \tag{4}$$

then we can define

$$\tau(X,Y) = \Pr\left[\text{csign}(X_i, Y_i, X_j, Y_j) = 1\right] - \Pr[\text{csign}(X_i, Y_i, X_j, Y_j) = -1], \tag{5}$$

where $\Pr(\cdot)$ is the probability of an event. In other words, $\tau(X,Y)$ is the difference between the probabilities of **concordance** (csign$(X_i, Y_i, X_j, Y_j) = 1$) and **discordance** (csign$(X_i, Y_i, X_j, Y_j) = -1$).

Returning to our variables (A_i, B_i), which are usually assumed to be sampled independently and identically from a common population of indexed individuals with a common bivariate distribution, we might view $\tau(A,B)$ as a measure of the concordance component of agreement and/or of the discordance component of disagreement. If the indexed population is a population of students, and (A_i, B_i) is the numbers of exam marks awarded by Lecturers A and B respectively, then $\tau(A,B)$ is the difference between the probability that they agree and the probability that they disagree, assuming that they are given 2 different exam scripts at random and are both asked which one is the best. Specifically, in the 176 students with total marks, the Kendall's τ_a between the A_i and the B_i is 0.708 (95% CI, 0.649 to 0.758; $P = 2.6 \times 10^{-36}$). So, with the 176 students in our sample, the two lecturers were 70.8 percent more likely to agree than to disagree. And, in the population of students at large, from which these 176 students were sampled, we are 95% confident that the two lecturers would be 64.9% to 75.8% more likely to agree than to disagree. (This confidence interval is asymmetric because it was computed using the Normalizing hyperbolic arctangent or Fisher's z transformation.) And the P–value shows that, in a fantasy scenario in which the lecturers were equally likely to agree or to disagree about pairs of students sampled randomly from the population at large, such a level of agreement in a sample would be *very* rare.

Returning to Figures 1 and 2, we note that the scatter plot of Figure 1 draws attention to concordance/discordance, because it has empty space in the upper left and lower right parts of the plot region, and therefore draws attention to the fact that students with a higher A_i from Lecturer A *usually* also have a higher B_i from Lecturer B. As we shall see, the Bland–Altman plot of Figure 2 draws attention to the other two components of agreement/disagreement.

2.1 Kendall's τ_a *versus* Pearson correlation

Statisticians are frequently heard to assert that the Pearson correlation coefficient "does not measure agreement". *However*, *if* a bivariate (X,Y) is sampled from a bivariate Normal distribution, or from any other bivariate distribution that can be transformed to bivariate Normal using a pair of monotonic transformations $g(X)$ and $h(Y)$ that may or may not be identity transformations, *then* the Pearson correlation coefficient between the variables (transformed if necessary) is equal to

$$\rho[g(X), h(Y)] = \sin\left[\frac{\pi}{2}\tau(X,Y)\right]. \tag{6}$$

(Note that we do not have to know the form of $g(\cdot)$ and $h(\cdot)$ in order for this to apply. Note, also, that this relation is monotonically increasing, and invertible, over the closed interval from -1 to 1.) This relation is known as **Greiner's relation**, and is discussed in Kendall (1949)[4], where it is pointed out that it also applies to a wide range of non–Normal bivariate distributions. Therefore, under a wide range of assumptions, the Pearson product–moment correlation $\rho(A,B)$ may measure at least one component of agreement/disagreement

between the A_i and the B_i, namely concordance/discordance. *However*, it does this in an indirect way, which cannot be interpreted simply as a difference between probabilities of agreement and disagreement.

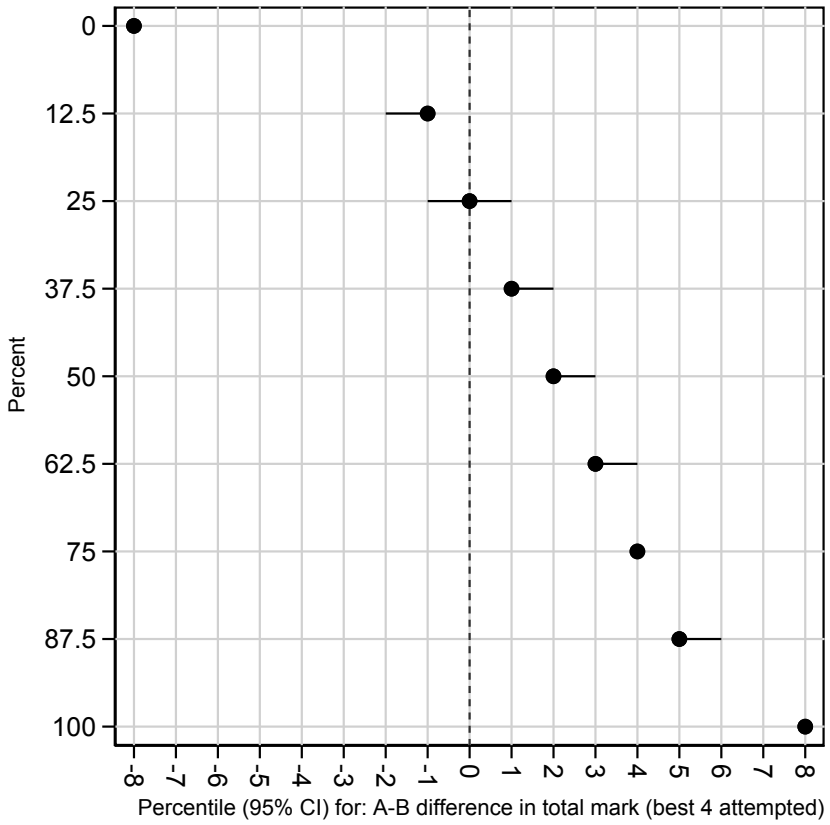
In our example dataset, the Pearson correlation between the marks awarded by the two lecturers, estimated using Greiner's relation to transform the estimate and confidence limits for Kendall's τ_a, was 0.896 (95% CI, 0.852 to 0.928). This is in good agreement with the directly–estimated Pearson correlation of 0.908. Under Greiner's relation, the Pearson correlation has a higher magnitude than the corresponding Kendall's τ_a. It is therefore important for the audience to understand the meaning of Kendall's τ_a as a difference between concordance and discordance probabilities. They should not become confused, just because they are accustomed to the higher–magnitude Pearson correlations and find the lower–magnitude Kendall's τ_a less impressive.

3 Bias parameters

The second component of disagreement is bias, which we commonly measure using a paired t–test to derive a confidence interval for the mean of the differences $A_i - B_i$. A possible alternative measure of bias might be percentile differences between the A_i and B_i. We can estimate the median (equal to the mean in a bivariate Normal model), and also other percentiles, to show how differences between measurements on the same subject vary within the population of subjects.

A useful Stata package for measuring percentiles is `rcentile` (Newson, 2014)[8], which allows adjustment of confidence intervals for clustering and/or weighting if necessary, and also saves the confidence intervals conveniently in a Stata matrix for the user to use. (In default, these confidence intervals for percentiles are calculated using the Normalizing and variance–stabilizing hyperbolic arctangent or Fisher's z transform on the mean signs of the differences between data values and percentiles.)

Figure 3: Percentile differences between marks awarded to the same candidate by Lecturers A and B.



In our example dataset, the percentile differences in marks awarded to the same candidate by Lecturers A and B are plotted against percents (with increments of 12.5% or 1/8) in Figure 3. They are also tabulated in Table 1. Note that, for most of the percentiles, the upper and/or lower 95% confidence limits are equal to the estimate. This is a consequence of the fact that the differences are all integers, implying that percentile differences and their confidence limits can only be integers or half-integers. However, this does not invalidate

Table 1: Percentile differences between marks awarded to the same candidate by Lecturers A and B.

Percent	*Percentile*	*(95%*	*CI)*
0	-8	(-8,	-8)
12.5	-1	(-2,	-1)
25	0	(-1,	1)
37.5	1	(1,	2)
50	2	(2,	3)
62.5	3	(3,	4)
75	4	(4,	4)
87.5	5	(5,	6)
100	8	(8,	8)

the confidence intervals, except in the case of Percentiles 0 and 100, which are the minimum and maximum, respectively, and which are not covered by the Central Limit Theorem. We see that, most of the time, Lecturer A is "Mr Nice", who allocates a higher mark than Lecturer B to the same student. *However*, Percentiles 12.5 and 25 indicate that, *sometimes*, Lecturer B is more generous. The paired t–test on these data gave a mean difference of 2.04 (95% CI, 1.62 to 2.46; $P = 9.3 \times 10^{-18}$). This tells us the positive *mean* difference, but does not tell us about the minority of exceptional negative differences.

3.1 Tests and confidence limits for the mean sign

Confidence intervals for percentiles do not always come with P–values. *However*, in this case, the parameter to test for a zero value is the mean sign, defined as $E[\text{sign}(A_i - B_i)]$. This is the parameter tested by the sign test. It can be estimated, with confidence limits and a P–value, using the `scsomersd` package in Stata, which can be downloaded from the Statistical Software Components (SSC) archive, and which requires two other SSC packages (`somersd` and `expgen`) in order to work.

To estimate the mean sign of the difference, assuming that the difference itself is stored in a variable named `dtotmark`, we type, in Stata,

```
scsomersd dtotmark 0, transf(z) tdist
```
(7)

and the mean sign is displayed with confidence limits and a P–value, once again using the hyperbolic arctangent or Fisher's z transformation. In our data, it is 0.534 (95% CI, 0.404 to 0.643; $P = 5.1 \times 10^{-11}$). This means that, in our sample, Lecturer A awarded the higher mark 53.4% more often than Lecturer B awarded the higher mark. And, in the population from which these students were sampled, the former event would happen 40.4% to 64.3% more often than the latter event. And the P–value shows that a mean sign of this magnitude would very rarely happen by chance, if both lecturers were equally likely to award the higher mark.

Looking at the Bland–Altman plot of Figure 2), we get the impression that there are more data points above the line of zero difference than below the line of zero difference. The mean sign, and the percentile differences, support this impression.

4 Scale differential parameters

Two methods may also disagree on the scale of the differences between values. In our example dataset, we might ask whether Lecturer A or Lecturer B awarded marks that differed more, generating a greater difference between the better exam scripts and the worse exam scripts.

A good estimate of this tendency, using rank methods, is

$$\tau[A - B, (A + B)/2] \;=\; \tau(A - B, A + B)\,, \tag{8}$$

which is the difference between the probabilities of concordance and discordance between the differences and the means of two (A_i, B_i) pairs. In a Bland–Altman plot, such as Figure 2, a positive (or negative) τ_a between mean and difference indicates a trend of increasing (or decreasing) differences with increasing means. A positive $\tau(A - B, A + B)$ indicates that a random pair of A_i usually differ by more than the corresponding pair of B_i, and a negative $\tau(A - B, A + B)$ indicates that the A_i usually differ by less than the corresponding B_i, at least in the absolute values of the differences.

4.1 Mathematical excursis

Our assertion that $\tau[A-B,(A+B)/2]$ measures scale differential needs to be proved mathematically. *However*, the proof is long–winded and can be ignored, and taken on trust, by readers who do not like equations, and who may prefer to move straight to the next Subsection and to see a confidence interval.

To prove our assertion, we assume that we are sampling two bivariate data points (A_i, B_i) and $A_j, B_j)$ independently from the same population. We define

$$\begin{array}{rl} \Delta_A = & |A_i - A_j|, \\ \Delta_B = & |B_i - B_j|, \\ \gamma_A = & \text{sign}(A_i - A_j), \\ \gamma_B = & \text{sign}(B_i - B_j), \end{array} \tag{9}$$

where $|\cdot|$ denotes the absolute value. We note that the differences between the differences between, and sums of, the A–values and B–values in the ith and jth bivariate pairs are given, respectively, by

$$\begin{array}{rl} (A_i - B_i) - (A_j - B_j) = & \gamma_A\Delta_A - \gamma_B\Delta_B, \\ (A_i + B_i) - (A_j + B_j) = & \gamma_A\Delta_A + \gamma_B\Delta_B. \end{array} \tag{10}$$

It follows that the τ_a between the differences $A_h - B_h$ and the sums $A_h + B_h$ is given by

$$\tau(A-B, A+B) \;=\; E\left[\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B)\right], \tag{11}$$

which is the expectation of a product of two factors, and which is $+1$ if both factors have the same nonzero value, -1 if both factors have different nonzero values, and zero otherwise.

When comparing the scales of variation of the A_h and of the B_h, we aim to compare the probabilities of the events $\Delta_A > \Delta_B$ and $\Delta_A < \Delta_B$, without forgetting that there is a third possible event $\Delta_A = \Delta_B$.

In the first possible event, in which $\Delta_A > \Delta_B$, we have

$$\begin{array}{rl} \text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B) = & \gamma_A, \\ \text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) = & \gamma_A, \end{array} \tag{12}$$

implying that the product of the above two factors is

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; \gamma_A^2 \;=\; 1, \tag{13}$$

because γ_A cannot be zero if $\Delta_A > \Delta_B \geq 0$.

Similarly, in the second possible event, in which $\Delta_A < \Delta_B$, we have

$$\begin{array}{rl} \text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B) = & -\gamma_B, \\ \text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) = & \gamma_B, \end{array} \tag{14}$$

implying that the product of the above two factors is

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; -\gamma_B^2 \;=\; -1, \tag{15}$$

because γ_B cannot be zero if $\Delta_B > \Delta_A \geq 0$.

In the third possible event, in which $\Delta_A = \Delta_B$, we have either $\Delta_A = \Delta_B = 0$ or $\Delta_A = \Delta_B > 0$. In the first instance $\Delta_A = \Delta_B = 0$, we have

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; 0 \times 0 \;=\; 0. \tag{16}$$

And, in the second instance $\Delta_A = \Delta_B > 0$, the signs γ_A and γ_B must both be nonzero, and in the set $\{-1, 1\}$. If the signs are the same, then we have $\gamma_A = \gamma_B$, implying that

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B) \;=\; \text{sign}\left[(\gamma_A - \gamma_A)\Delta_B\right] \;=\; 0. \tag{17}$$

And, if the signs are different, then we have $\gamma_A = -\gamma_B$, implying that

$$\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; \text{sign}\left[(\gamma_A - \gamma_A)\Delta_B\right] \;=\; 0. \tag{18}$$

Therefore, if $\Delta_A = \Delta_B$, then, by (16), (17) and (18), we must have

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; 0. \tag{19}$$

It follows from (13), (15) and(19) that

$$\text{sign}(\gamma_A\Delta_A - \gamma_B\Delta_B)\text{sign}(\gamma_A\Delta_A + \gamma_B\Delta_B) \;=\; \begin{cases} 1 & , \Delta_A > \Delta_B, \\ -1 & , \Delta_A < \Delta_B, \\ 0 & , \Delta_A = \Delta_B, \end{cases} \tag{20}$$

implying that (11) can be restated as

$$\tau(A-B, A+B) \;=\; E\left[\text{sign}(\Delta_A - \Delta_B)\right] \;=\; \Pr(\Delta_A > \Delta_B) - \Pr(\Delta_A < \Delta_B). \tag{21}$$

In other words, $\tau[A-B, (A+B)/2] = \tau(A-B, A+B)$ is the difference between two probabilities, namely the probability that the absolute difference between two random A–values is greater than the absolute difference between the corresponding B–values and the probability that the absolute difference between two random A–values is less than the absolute difference between the corresponding B–values.

4.2 Mean–difference τ_a in the example dataset

Returning to our example dataset, we find that the Kendall's τ_a between the means $(A_i + B_i)/2$ and the differences $A_i - B_i$ is 0.266 (95% CI, 0.169 to 0.358; $P = 3.8 \times 10^{-07}$). This means that, in our sample of students, if we choose two students at random to be marked by Lecturer A and by Lecturer B, then it is 26.6% more likely that the difference between the better script and the worse script will be greater according to Lecturer A than according to Lecturer B than that this difference will be greater according to Lecturer B than according to Lecturer A. And, in the population of students at large, we are 95% confident that it would be 16.9% to 35.8% more likely. And the P–value indicates that this scale difference is not likely to be generated by sampling error, in a fantasy scenario where both lecturers grade students on the same scale. This may be because the more experienced and confident Lecturer A was more discriminating than the less–experienced and caffeine–overloaded Lecturer B. Once again, the confidence interval and the P–value were calculated using the hyperbolic arctangent or Fisher's z transformation.

Looking at the Bland–Altman plot in Figure 2, we seem to see (by eye) that larger means are *usually* found with larger differences. This impression is supported by the statistics.

4.3 Greiner's relation for means and differences

The Greiner relation (6) may still apply for Kendall's τ_a between the means $(A_i + B_i)/2$ and the differences $A_i - B_i$, especially if the A_i and B_i have a bivariate Normal joint distribution, in which case so will the $A_i - B_i$ and $(A_i + B_i)/2$. In this case, the Pearson correlation between the means and the differences is given (after some algebraic manipulation) by

$$\rho\left[A-B, (A+B)/2\right] \;=\; \rho(A-B, A+B) \;=\; \frac{\text{Var}(A) - \text{Var}(B)}{\sqrt{\left[\text{Var}(A)\right]^2 + \left[\text{Var}(B)\right]^2 + 2\text{Var}(A)\text{Var}(B)\left[1 - 2\rho(A,B)^2\right]}}, \tag{22}$$

where $\text{Var}(\cdot)$ denotes variance. By the Schwarz inequality, the denominator must be positive if the variances are positive. This implies that (22) is positive, negative or zero if $\text{Var}(A) - \text{Var}(B)$ is positive, negative or zero, respectively. It is therefore interpretable as a measure of scale differential between the variances of the A_i and of the B_i, at least in principle. *However*, it does not have an interpretation as a difference between probabilities, as $\tau(A-B, A+B)$ does.

In our example dataset, the Pearson correlation derived by transforming $\tau[A-B, (A+B)/2]$ using Greiner's relation is 0.406 (95% CI, 0.263 to 0.533; $P = 3.8 \times 10^{-07}$). This seems to imply that the marks of Lecturer A are more variable than the marks of Lecturer B, and agrees well with the corresponding directly–calculated Pearson correlation of 0.424. However, I cannot immediately think of any further interpretation.

5 Log–scale Bland–Altman plots for ratios and geometric means of positive–valued measures

Sometimes, especially in the analysis of DNA microarray data, log–scale Bland–Altman plots are produced for strictly positive–valued measures, where the ratio of measurements A_i/B_i on the vertical axis is plotted against the geometric mean (GM) of measurements A_i and B_i (equal to $\sqrt{A_i B_i}$) on the horizontal axis, both on a log scale (usually binary). These are often known by other names, such as "MA plots" or "RA plots"[3]. The rank parameters for these plots are defined by substituting the $\ln A_i$ and $\ln B_i$ in the formulas of the previous sections, and exponentiating differences between logs to give ratios.

Figure 4: Binary log–scale scatter plot of marks awarded by Lecturers A and B.

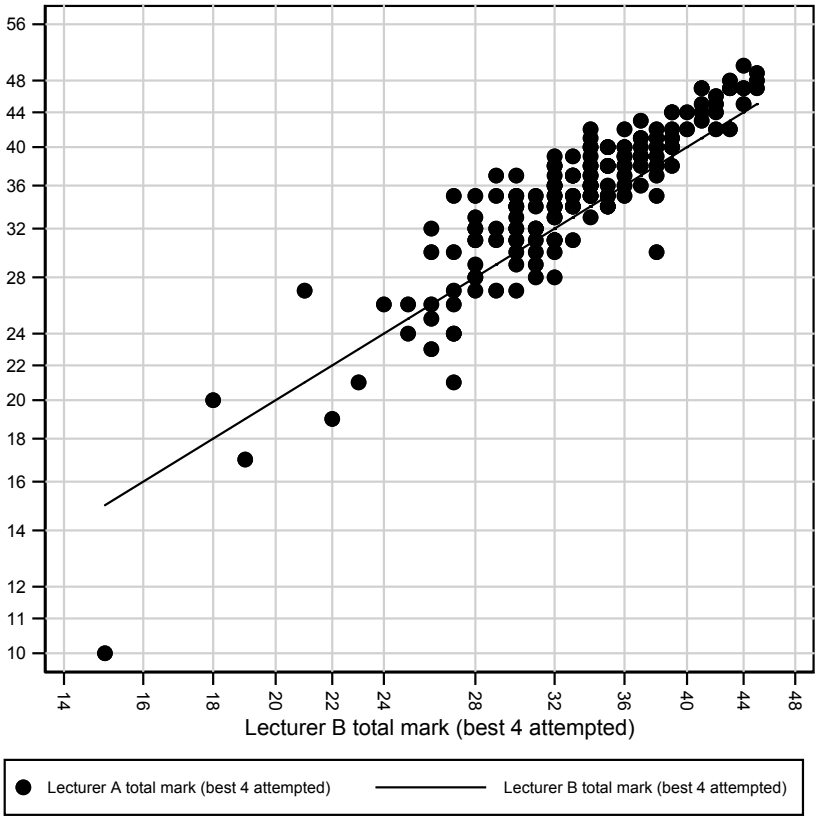
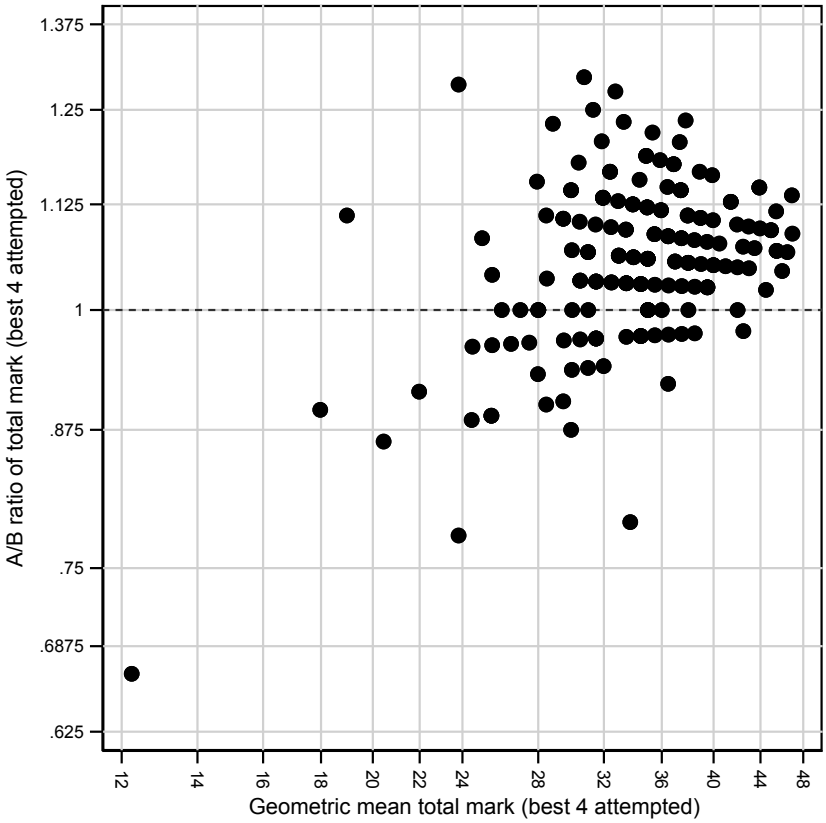


Figure 5: Binary log–scale Bland–Altman plot of marks awarded by Lecturers A and B.



For our example dataset, the binary log–scale scatter plot is given in Figure 4, and the binary log–scale Bland–Altman plot is given in Figure 5. Note that, this time, the vertical–axis reference line of the Bland–Altman plot is at the "null ratio" of 1, expected if both lecturers are equally generous.

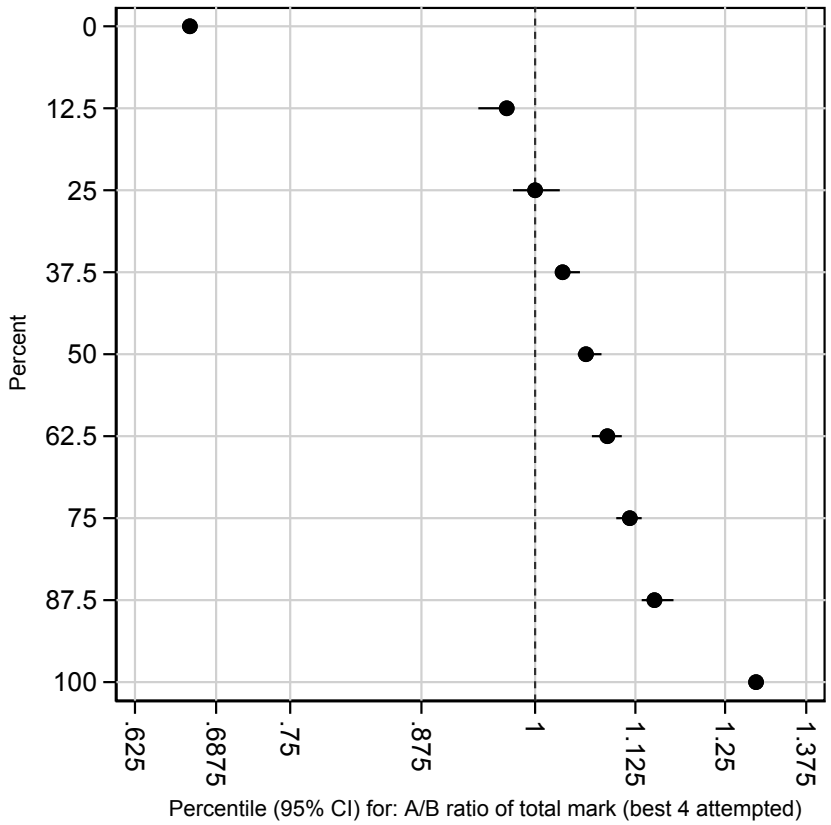
The discordance parameter is identical for log–scale and linear–scale data, because $\tau(\ln A, \ln B) = \tau(A, B)$ for positive–valued variables A and B. So, once again, the Kendall's τ_a between the marks of the 2 lecturers is 0.708 (95% CI, 0.649 to 0.758; $P = 2.6 \times 10^{-36}$). And the Pearson correlation from Greiner's relation is once again 0.896 (95% CI, 0.852 to 0.928; $P = 2.6 \times 10^{-36}$). This is in good agreement with the directly–estimated Pearson correlation of the logs, which is 0.905.

The bias parameters this time are the percentiles of the ratios of marks awarded by Lecturer A and Lecturer B. They are plotted in Figure 6 and tabulated in Table 2. Again, we see that Lecturer A is usually (but not always) more generous than Lecturer B. The mean sign, with confidence limits and a P–value, is this time the mean sign of the differences between the A_i/B_i ratios and 1, and is estimated using the command

```
scsomersd rtotmark 1, transf(z) tdist
```
(23)

assuming that the ratios A_i/B_i are stored in the variable `rtotmark`. This mean sign is, of course, identical to the mean sign of the differences between the $A_i - B_i$ ratios and zero, namely 0.534 (95% CI, 0.404 to 0.643; $P = 5.1 \times 10^{-11}$). The paired t–test between the $\ln A_i$ and the $\ln B_i$ is this time computed to give an exponentiated confidence interval for the GM ratio, which is 1.055 (95% CI, 1.040 to 1.069; $P = 1.1 \times 10^{-12}$). This is in good agreement with the median ratio.

Figure 6: Percentile ratios between marks awarded to the same candidate by Lecturers A and B.



The scale differential parameter this time is the Kendall's τ_a between the A_i/B_i ratios and the geometric means. This time, it is equal to 0.163 (95% CI, 0.054 to 0.269; $P = .0038$). So, if a random pair of exam scripts is marked by Lecturer A and Lecturer B, then the ratio between the higher and lower of the 2 marks awarded by the same lecturer is 16.3% more likely to be greater when awarded by Lecturer A than when awarded by Lecturer B than it is to be greater when awarded by Lecturer B than when awarded by Lecturer A. And, in the population at large, the difference between the two probabilities is probably between 5.4% to 26.9%. Therefore, Lecturer A seems to be more discriminating than Lecturer B in relative terms, as well as in absolute terms. However, the difference in relative discrimination seems to be less than the difference in absolute discrimination, probably because Lecturer A is typically more generous at marking

Table 2: Percentile ratios between marks awarded to the same candidate by Lecturers A and B.

Percent	*Percentile*	*(95%*	*CI)*
0	0.667	(0.667,	0.667)
12.5	0.967	(0.935,	0.974)
25	1.000	(0.974,	1.029)
37.5	1.033	(1.028,	1.054)
50	1.062	(1.051,	1.081)
62.5	1.089	(1.069,	1.107)
75	1.118	(1.100,	1.133)
87.5	1.150	(1.133,	1.176)
100	1.296	(1.296,	1.296)

both scripts in a pair. The Pearson correlation between ratios and GMs estimated using Greiner's relation is 0.254 (95% CI, 0.085 to 0.409; $P = .0038$). This is not in very good agreement with the directly–estimated Pearson correlation between the log ratios and the log GMs, which is 0.433. This is probably because the log is not really a sensible Normalizing transformation to use with these data, which seem (if anything) to be negatively skewed, with an outlier in the bottom–left corner of Figures 4 and 5. This distribution–sensitivity of the regression–based parameter is probably another good reason for preferring rank parameters to regression parameters when summarizing Bland–Altman plots.

References

[1] Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 1983; **32(3)**: 307-317.

[2] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i(8476)**: 307–310.

[3] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; **12(1)**: 111–139.

[4] Kendall MG. Rank and product–moment correlation. *Biometrika* 1949; **36(1/2)**: 177-193.

[5] Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45(1)**: 255–268.

[6] Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal* 2006; **6(3)**: 309–334. Download from *http://www.stata-journal.com/article.html?article=snp15_6*

[7] Newson R. Confidence intervals for rank statistics: Percentile slopes, differences, and ratios. *The Stata Journal* 2006; **6(4)**: 497–520. Download from *http://www.stata-journal.com/article.html?article=snp15_7*

[8] Newson RB. Easy–to–use packages for estimating rank and spline parameters. Presented at the 20th UK Stata User Meeting, 11–12 September, 2014. Download from *https://ideas.repec.org/p/boc/usug14/01.html*

[9] StataCorp. *Stata: Release 15. Statistical Software.* College Station, TX: StataCorp LLC; 2017.

[10] van Belle G. Statistical Rules of Thumb. Second Edition. Hoboken, NJ: John Wiley & Sons, Inc.; 2008.